31250 Introduction to Data Analytics – Assignment 2

[1. Initial Data Exploration 2](#_Toc146912602)

[1.1 Attribute Types 2](#_Toc146912603)

[1.2 The Summarising Properties for the Attributes 5](#_Toc146912604)

[1.2.1 gender 5](#_Toc146912605)

[1.2.2 age 6](#_Toc146912606)

[1.2.3 LOSdays 8](#_Toc146912607)

[1.2.4 admit\_location 9](#_Toc146912608)

[1.2.5 AdmitDiagnosis 11](#_Toc146912609)

[1.2.6 Insurance 12](#_Toc146912610)

[1.2.7 NumCallouts 13](#_Toc146912611)

[1.2.8 NumDiagnosis 15](#_Toc146912612)

[1.2.9 NumProcs 16](#_Toc146912613)

[1.2.10 AdmitProcedure 17](#_Toc146912614)

[1.2.11 NumCPTevents 18](#_Toc146912615)

[1.2.12 NumInput 20](#_Toc146912616)

[1.2.13 NumLabs 21](#_Toc146912617)

[1.2.14 NumMicroLabs 22](#_Toc146912618)

[1.2.15 NumOutput 23](#_Toc146912619)

[1.2.16 NumTransfers 25](#_Toc146912620)

[1.2.17 NumChartEvents 26](#_Toc146912621)

[1.2.18 ExpiredHospital 27](#_Toc146912622)

[1.2.19 TotalNumInteract 29](#_Toc146912623)

[1.2.20 Marital status 30](#_Toc146912624)

[1.3 Exploration 32](#_Toc146912625)

[1.3.1 Outliers 32](#_Toc146912626)

[1.3.2 Interesting attributes 34](#_Toc146912627)

[1.3.3 Clusters 39](#_Toc146912628)

[2. Data Preprocessing 39](#_Toc146912629)

[2.1 Binning of age 39](#_Toc146912630)

[2.1.1 Equiwidth Binning 40](#_Toc146912631)

[2.1.2 Equidepth binning 42](#_Toc146912632)

[2.2 Normalisation of NumLabs 43](#_Toc146912633)

[2.2.1 Min-max normalisation 43](#_Toc146912634)

[2.2.2 Z-score normalisation 44](#_Toc146912635)

[2.3 Discretisation of LOSdays 44](#_Toc146912636)

[2.4 Binarisation of marital status 45](#_Toc146912637)

[3. Summary 47](#_Toc146912638)

# 1. Initial Data Exploration

## 1.1 Attribute Types

**Attribute Name:** gender **Attribute Type:** Nominal **Justification:** The attribute type of gender is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** age **Attribute Type:** Ratio **Justification:** The attribute type of age is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero years.

**Attribute Name:** LOSdays **Attribute Type:** Ratio **Justification:** The attribute type of LOSdays is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero days.

**Attribute Name:** admit\_location **Attribute Type:** Nominal **Justification:** The attribute type of admit\_location is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** AdmitDiagnosis **Attribute Type:** Nominal **Justification:** The attribute type of AdmitDiagnosis is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** Insurance **Attribute Type:** Nominal **Justification:** The attribute type of Insurance is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** NumCallouts **Attribute Type:** Ratio **Justification:** The attribute type of NumCallouts is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero callouts.

**Attribute Name:** NumDiagnosis **Attribute Type:** Ratio **Justification:** The attribute type of NumDiagnosis is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero diagnoses.

**Attribute Name:** NumProcs **Attribute Type:** Ratio **Justification:** The attribute type of numProcs is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero procedures.

**Attribute Name:** AdmitProcedure **Attribute Type:** Nominal **Justification:** The attribute type of AdmitProcedure is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** NumCPTevents **Attribute Type:** Ratio **Justification:** The attribute type of NumCPTevents is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero events.

**Attribute Name:** NumInput **Attribute Type:** Ratio **Justification:** The attribute type of NumInput is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero inputs.

**Attribute Name:** NumLabs **Attribute Type:** Ratio **Justification:** The attribute type of NumLabs is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero labs.

**Attribute Name:** NumMicroLabs **Attribute Type:** Ratio **Justification:** The attribute type of NumMicroLabs is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero micro labs.

**Attribute Name:** NumOutput **Attribute Type:** Ratio **Justification:** The attribute type of NumOutput is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero outputs.

**Attribute Name:** NumTransfers **Attribute Type:** Ratio **Justification:** The attribute type of NumTransfers is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero transfers.

**Attribute Name:** NumChartEvents **Attribute Type:** Ratio **Justification:** The attribute type of NumChartEvents is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero events.

**Attribute Name:** ExpiredHospital **Attribute Type:** Nominal **Justification:** The attribute type of ExpiredHospital is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

**Attribute Name:** TotalNumInteract **Attribute Type:** Ratio **Justification:** The attribute type of TotalNumInteract is ‘Ratio” as it represents units that have numeric values that can be ordered as the exact differences between the values can be calculated. Additionally, it has an origin point of zero interactions.

**Attribute Name:** marital status **Attribute Type:** Nominal **Justification:** The attribute type of marital status is ‘Nominal’ as it cannot be ordered, it represents discrete units and is used to label variables, that have no quantitative value.

## 1.2 The Summarising Properties for the Attributes

### 1.2.1 gender

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | M |
| Unique Values | 2 |
| Missing Values | 0 |

As gender is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique Values, Missing Values and Frequency statistics. The below pie chart (See Figure 1) visualizes the frequency statistics and indicates the values (F, M) that support the gender attribute. ‘F’ represents female patients and ‘M’ represents male patients. From the row count we find that 43.58% of patients are female and 56.42% of patients are male. From the frequency bar chart (See Figure 2) we see men slightly outnumber women with 1331 entries to 1028 entries, hence the mode is ‘M’. A possible explanation for this finding is that men typically engage in more dangerous activities.

*Figure 1 Pie Chart for gender:*

A graph of a pie chart

Description automatically generated

*Figure 2 Bar Chart for gender:*

A screenshot of a graph

Description automatically generated

### 1.2.2 age

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 52.459 |
| Minimum Value | 0 |
| Maximum Value | 88 |
| Standard Deviation | 26.349 |
| Variance | 694.246 |
| Range | 88 |
| Mode | 0 |
| Skewness | -0.861 |
| Unique Values | 74 |
| 25% Quantile | 42 |
| Median | 58 |
| 75% Quantile | 73 |
| Sum | 123,750 |
| Kurtosis | -0.861 |
| Missing Values | 0 |

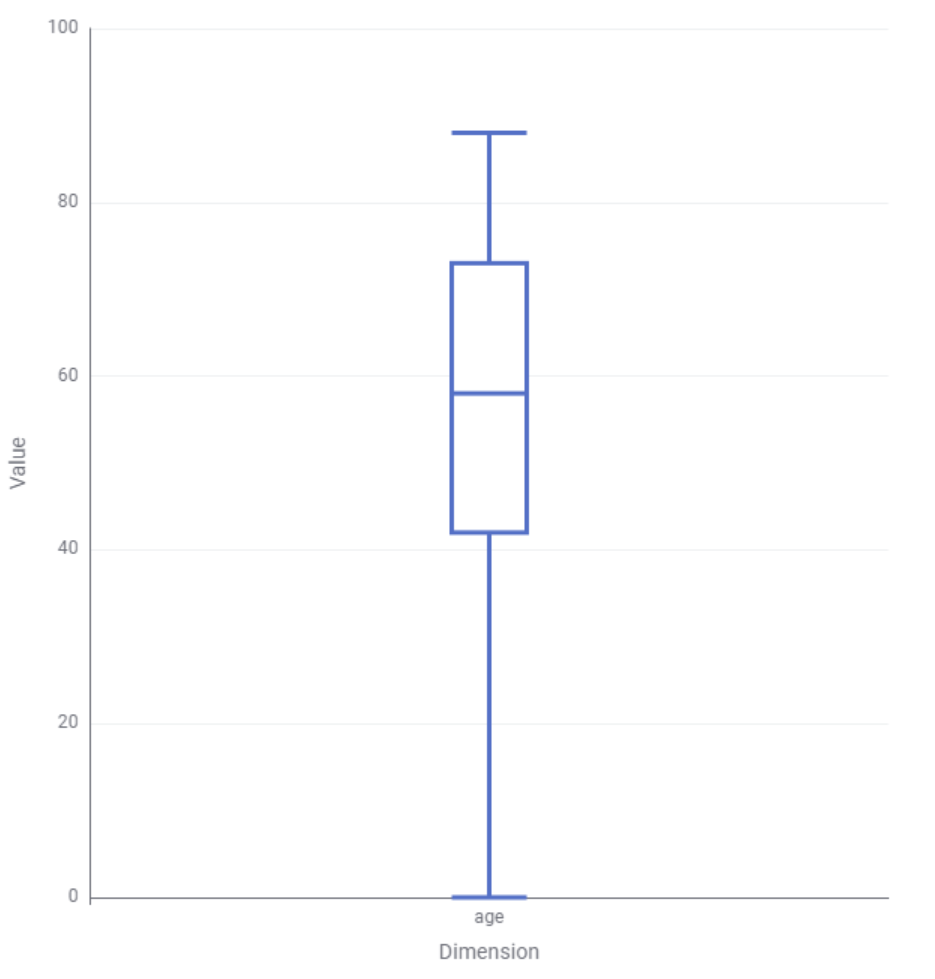
As age is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 3) visualizes the row count frequency statistics of patient ages. Binning has been used to best capture the shape of the distribution. An age of 0 is by far the most common (having over 50% more entries than the next most common age bracket). This could be explained by the hospital dealing with a large number of child births. From the quantiles we also find that 50% of patients are between 41 and 75 years old. This results in a strong negative skew to the age distribution as visualised by the box plot in figure 4. The variance is notably higher than the range which indicates a broad spread of data points.

*Figure 3 Histogram for age:*

A graph of a graph

Description automatically generated with medium confidence

*Figure 4 Box Plot for age:*



### 1.2.3 LOSdays

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 9.967 |
| Minimum Value | 0 |
| Maximum Value | 124.63 |
| Standard Deviation | 11.55 |
| Variance | 133.396 |
| Range | 124.63 |
| Mode | 3.92 |
| Skewness | 3.627 |
| Unique Values | 650 |
| 25% Quantile | 3.63 |
| Median | 6.33 |
| 75% Quantile | 11.92 |
| Sum | 23,512.12 |
| Kurtosis | 19.907 |
| Missing Values | 0 |

As LOSDays is a ratio attribute that describes the length of stay at the hospital, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 5) visualizes the row count frequency statistics for the length in days of patient stays at the hospital. Binning has been used to best capture the shape of the distribution. The most common (mode) length of stay is 3.92, this can be used to predict patient turnover and hospital capacity. Additionally, Figure 5 depicts the strong positive skew of the data which indicates that shorter stays are most common. Similarly, the below box plot (See Figure 6) clearly depicts that the majority of patients stay for less than 10 days despite the range being over 12 times the size this interval.

*Figure 5 Histogram for LOSdays:*

A graph of a person with a beard

Description automatically generated with medium confidence

*Figure 6 Box Plot for LOSdays:*

A diagram of a graph

Description automatically generated

### 1.2.4 admit\_location

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | EMERGENCY ROOM ADMIT |
| Unique Values | 9 |
| Missing Values | 0 |

As admit\_location is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique Values, Missing values and Frequency statistics. The below pie chart (See Figure 7) visualizes the row count frequency statistics and indicates the 9 values that support the admit\_location attribute (CLINIC REFERRAL/PREMATURE, EMERGENY ROOM ADMIT, \*\*INFO NOT ABAILABLE\*\* HMO REFERRAL/STICK, PHYS REFERRAL/NORMAL DELI, TRANFSER FROM HOSP/EXTRAM, TRANSFER FROM OTHER HEALT, TRANSFER FROM SKILLED NUR, TRSF WITHIN THIS FACILITY). The most common (mode) admission location is EMERGENCY ROOM ADMIT which makes up 37.39% of all cases. In consecutive order of frequency (See Figure 8) PHYS REFERRAL/NORMAL DELI, CLINIC REFERRAL/PREMATURE and TRANSFER FROM HOSP/EXTRAM are the next most frequent admit locations. From Figure 7 all other admit locations only make up the remaining 1.27% of patients (A threshold of .56% was used to display this). Although there are 0 missing values in the data, attributes labelled with \*\*INFO NOT AVIALABLE\* can be considered tangibly as missing values.

*Figure 7 Pie Chart for admit\_location:*

A colorful pie chart with text

Description automatically generated

*Figure 8 Histogram for admit\_location:*

A graph of a bar graph

Description automatically generated with medium confidence

### 1.2.5 AdmitDiagnosis

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | NEWBORN |
| Unique Values | 1063 |
| Missing Values | 5 |

As AdmitDiagnosis is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique values, Missing values and frequency statistics. The below pie chart (See Figure 9) visualizes the row count frequency statistics and indicates the values that support the AdmitDiagnosis attribute. The most common (mode) reason for patient admission is ‘NEWBORN’ (i.e., children born at the hospital) which makes up 13.69% or 323 entries of all cases. There are 1064 other attribute values, where the next most common of which (PNEUMONIA) only makes up approximately 1/5 that of NEWBORN cases. Hence, no other admit diagnosis is as dominant as newborns. Consequently, attributes with a frequency below the threshold of .4% are placed into the Other category for maximized readability of Figure 9. This category makes up 60.3% of AdmitDiagnosis values, hence there is a large range of similarly frequent values for AdmitDiagnosis.

*Figure 9 Pie Chart for AdmitDiagnosis:*

A colorful pie chart with text

Description automatically generated

### 1.2.6 Insurance

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | Medicare |
| Unique Values | 5 |
| Missing Values | 0 |

As Insurance is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique values, Missing values and frequency statistics. The below pie chart (See Figure 10) and bar chart (See Figure 11) visualizes the row count frequency statistics and indicates the values (Medicare, Private, Medicaid, Government, Self Pay) that support the Insurance attribute. The most common (mode) type of insurance is Medicare which makes up 46.97% or 1108 entries of all cases. ‘Private’ insurance closely follows medica rein frequency, making up 38.58% of the rows. As Seen in Figure 10 and 11, even combined the remaining Insurance values do not make up a significant portion of the dataset relative to Medicare and Private.

*Figure 10 Pie Chart for Insurance:*

*A pie chart with different colored sections

Description automatically generated*

*Figure 11 Bar Chart for Insurance:*

A graph of blue squares

Description automatically generated

### 1.2.7 NumCallouts

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 0.105 |
| Minimum Value | 0 |
| Maximum Value | 2 |
| Standard Deviation | 0.166 |
| Variance | 0.028 |
| Range | 2 |
| Mode | 0 |
| Skewness | 3.521 |
| Unique Values | 79 |
| 25% Quantile | 0 |
| Median | 0 |
| 75% Quantile | 0.17 |
| Sum | 248.26 |
| Kurtosis | 23.501 |
| Missing Values | 0 |

As NumCallouts is a ratio attribute, quantitative summary statistics, of location and spread have been included. The below histogram (See Figure 12) visualizes the row count frequency statistics for the number of callouts. Binning has been used to best capture the shape of the distribution. The most common (mode) number of callouts is 0. Additionally, Figure 12 depicts the strong positive skew of the data which indicates that a lower number of callouts is most common. Similarly, the blow box plot (See Figure 13) clearly depicts that the majority (75%) of patients lie below 0.17 callouts. The dominance of 0 callouts is highlighted by the 1st and 2nd quartile both being 0.

*Figure 12 Histogram for NumCallouts:*

A graph with a number of squares

Description automatically generated with medium confidence

*Figure 13 Box Plot for NumCallouts:*

A diagram of a graph

Description automatically generated

### 1.2.8 NumDiagnosis

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 2.638 |
| Minimum Value | 0 |
| Maximum Value | 150 |
| Standard Deviation | 6.879 |
| Variance | 47.327 |
| Range | 150 |
| Mode | 1 |
| Skewness | 12.296 |
| Unique Values | 570 |
| 25% Quantile | .83 |
| Median | 1.4 |
| 75% Quantile | 2.5 |
| Sum | 6,222.57 |
| Kurtosis | 197.482 |
| Missing Values | 0 |

As NumDiagnosis is a ratio attribute, quantitative summary statistics, of location and spread have been included. The below histogram (See Figure 14) visualizes the row count frequency statistics for the number of callouts. Binning has been used to best capture the shape of the distribution. The most common (mode) number of callouts is 1. Additionally, Figure 14 depicts the strong positive skew of the data which indicates that a lower number of diagnosis is most common. Figure 14 has an upper bound approximately where outliers begin as seen in the box plot of Figure 15 so that readability of frequency is improved for the other values.

*Figure 14 Histogram for NumDiagnosis*

A graph of a number of bars

Description automatically generated with medium confidence

Figure 15 Box Plot for NumDiagnosis

A graph with numbers and dots

Description automatically generated

### 1.2.9 NumProcs

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | .712 |
| Minimum Value | 0 |
| Maximum Value | 100 |
| Standard Deviation | 2.417 |
| Variance | 5.84 |
| Range | 254 |
| Mode | 0 |
| Skewness | 30.544 |
| Unique Values | 254 |
| 25% Quantile | .21 |
| Median | .4 |
| 75% Quantile | .69 |
| Sum | 1,678.58 |
| Kurtosis | 1,215.275 |
| Missing Values | 0 |

As NumProcs is a ratio attribute, quantitative summary statistics, of location and spread have been included. The below histogram (See Figure 16) visualizes the row count frequency statistics for the number of callouts. Binning has been used to best capture the shape of the distribution. The most common (mode) number of procedures is 0. Additionally, Figure 16 depicts the strong positive skew of the data which indicates that a lower number of diagnosis is most common. Figure 16 has an upper bound that increases readability of frequency for the other values.

*Figure 16 Histogram for NumProcs:*

A graph of a graph

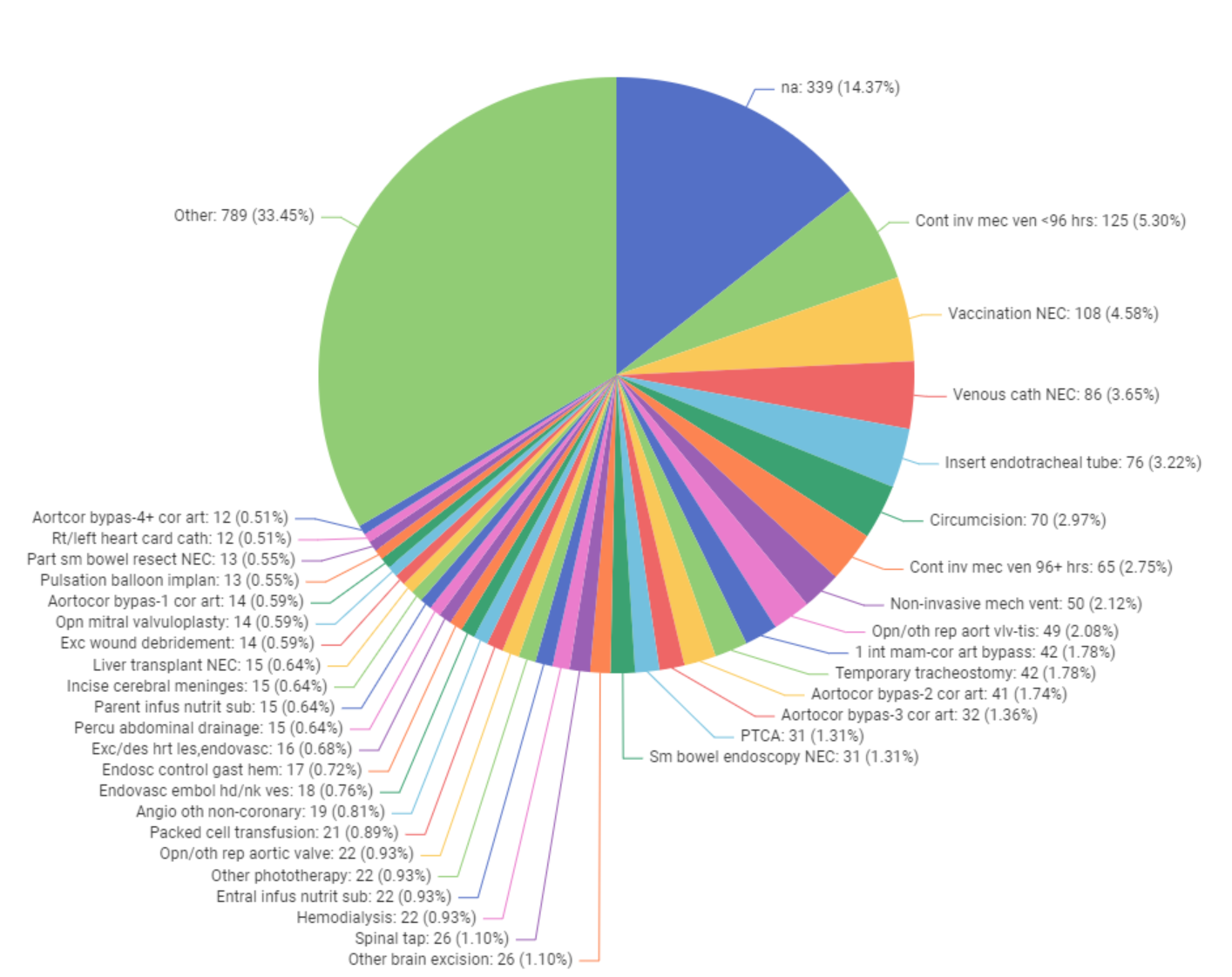
Description automatically generated with medium confidence

### 1.2.10 AdmitProcedure

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | EMERGENCY ROOM ADMIT |
| Unique Values | 9 |
| Missing Values | 0 |

As AdmitProcedure is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique values, Missing values and frequency statistics. The blow pie chart (See Figure 17) visualizes the row count frequency statistics and indicates the values that support the AdmitProcedure attribute. The most common (mode) admit procedure was ‘na’ i.e. not applicable, which makes up 14.37% or 339 entries of all cases. There are 380 other attribute values, where the next most common of which is ‘Cont inv mec ven < 96hrs’ at 5.3%. Hence no other AdmitProcedure is as dominant as ‘na’. Consequently, attributes with a frequency below the threshold of .5% are placed into the Other category for maximized readability of Figure 17. This category makes up 33.45% of AdmitProcedure values, hence there is a large range of similarly frequent values for AdmitDiagnosis.

*Figure 17 Pie Chart for AdmitProcedure*



### 1.2.11 NumCPTevents

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 1.067 |
| Minimum Value | 0 |
| Maximum Value | 23.53 |
| Standard Deviation | 1.293 |
| Variance | 1.673 |
| Range | 23.53 |
| Mode | 0 |
| Skewness | 6.942 |
| Unique Values | 342 |
| 25% Quantile | 0.04 |
| Median | 1 |
| 75% Quantile | 1.57 |
| Sum | 2,517.02 |
| Kurtosis | 95.656 |
| Missing Values | 0 |

As NumCPTevents is a ratio attribute, quantitative summary statistics, of location and spread have been included. The below histogram (See Figure 18) visualizes the row count frequency statistics for the number of CPT events. Binning has been used to best capture the shape of the distribution. The most common (mode) number of events is 0. Additionally, Figure 18 depicts the strong positive skew of the data which indicates that a lower number of events is most common. Figure 19 is a histogram with an upper bound of where outliers approximately begin as seen in the box plot of Figure 19 so that readability of frequency is improved for the other values. From Figure 19 we also see that 75% of values lie between 0 and 1.57.

*Figure 18 Histogram for NumCPTevents:*

A graph of a person standing in front of a white background

Description automatically generated

*Figure 19 Box Plot for NumCPTevents:*

A white background with black dots and a rectangular object

Description automatically generated

### 1.2.12 NumInput

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 28.151 |
| Minimum Value | 0 |
| Maximum Value | 614.4 |
| Standard Deviation | 45.938 |
| Variance | 2,110.34 |
| Range | 614.4 |
| Mode | 0 |
| Skewness | 4.674 |
| Unique Values | 32.926 |
| 25% Quantile | 4.58 |
| Median | 13.41 |
| 75% Quantile | 32.26 |
| Sum | 66,407.58 |
| Kurtosis | 32.926 |
| Missing Values | 0 |

As NumInputs is a ratio attribute, quantitative summary statistics of location and spread have been included. An input in a medical context refers to the amount of fluids, medications or nutrients given to a patient. The below histogram (See Figure 20) visualizes the row count frequency statistics of patient ages. Binning has been used to best capture the shape of the distribution. Figure 20 is histogram with an upper bound where outliers approximately begin as seen in the box plot of Figure 21 so that readability of frequency is improved for the other values. 0 is by far the most common number of inputs (having over 50% more entries than the next most common fluid quantity). From the quantiles we also find that 75% of values lie between 0 and 32.26. This results in a strong positive skew to the age distribution as visualized by the box plot in figure 21.

*Figure 20 Histogram for NumInputs*

A graph of a person with a bar graph

Description automatically generated with medium confidence

*Figure 21 Box Plot for NumInput:*

A black and white image of a line

Description automatically generated

### 1.2.13 NumLabs

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 47.871 |
| Minimum Value | 0 |
| Maximum Value | 5,175 |
| Standard Deviation | 120.545 |
| Variance | 14,531.095 |
| Range | 5175 |
| Mode | 0 |
| Skewness | 33.555 |
| Unique Values | 1919 |
| 25% Quantile | 26.99 |
| Median | 38.48 |
| 75% Quantile | 50.61 |
| Sum | 112,928.18 |
| Kurtosis | 1,393.885 |

As NumLabs is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 22) visualizes the row count frequency statistics for the number of labs. Binning has been used to best capture the shape of the distribution. The most common (mode) number of labs is 0. Additionally, Figure 22 depicts the positive skew of the data which indicates that a lower number of callouts is most common. Figure 22 is a histogram an upper bound so that readability of frequency is improved for the other values. Interestingly, Figure 22 seems to follow a more normal distribution which indicates that a nonzero number of labs centered on the median 38.48 is also to be expected.

*Figure 22 Histogram for NumLabs:*

A graph of a number of bars

Description automatically generated with medium confidence

### 1.2.14 NumMicroLabs

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 1.359 |
| Minimum Value | 0 |
| Maximum Value | 347.18 |
| Standard Deviation | 8.25 |
| Variance | 68.068 |
| Range | 347.18 |
| Mode | 0 |
| Skewness | 35.298 |
| Unique Values | 436 |
| 25% Quantile | .15 |
| Median | .52 |
| 75% Quantile | 1.37 |
| Sum | 3,206.43 |
| Kurtosis | 1393.297 |
| Missing Values | 0 |

As NumMicroLabs is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 23) visualizes the row count frequency statistics for the number of labs. Binning has been used to best capture the shape of the distribution, however due to the extreme outliers a readable histogram cannot be created without modification. The most common (mode) number of labs is 0. Additionally, Figure 23 depicts the strong positive skew of the data which indicates that a lower number of callouts is most common. Figure 23 is a histogram with an upper bound so that readability of frequency is improved for the other values.

*Figure 23 Histogram for NumMicroLabs:*

*A graph with a bar graph

Description automatically generated with medium confidence*

### 1.2.15 NumOutput

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 6.898 |
| Minimum Value | 0 |
| Maximum Value | 62.4 |
| Standard Deviation | 6.58 |
| Variance | 43.294 |
| Range | 62.4 |
| Mode | 0 |
| Skewness | 1.591 |
| Unique Values | 1216 |
| 25% Quantile | 1.85 |
| Median | 5.36 |
| 75% Quantile | 10.18 |
| Sum | 16272.79 |
| Kurtosis | 4.859 |
| Missing Values | 0 |

As NumOutput is a ratio attribute, quantitative summary statistics, of location and spread have been included. The below histogram (See Figure 24) visualizes the row count frequency statistics for the number of outputs (Where outputs refer removal of fluids etc from a patient). Binning has been used to best capture the shape of the distribution. The most common (mode) number of outputs is 0. Additionally, Figure 24 depicts the strong positive skew of the data which indicates that a lower number of outputs is most common. From Figure 25 we visualize the quantiles to see that 75% of values lie between 0 and 10.18 outputs.

*Figure 24 Histogram for NumOutput:*

A graph of a person

Description automatically generated with medium confidence

*Figure 25 Box Plot for NumOutput:*

A white background with black lines

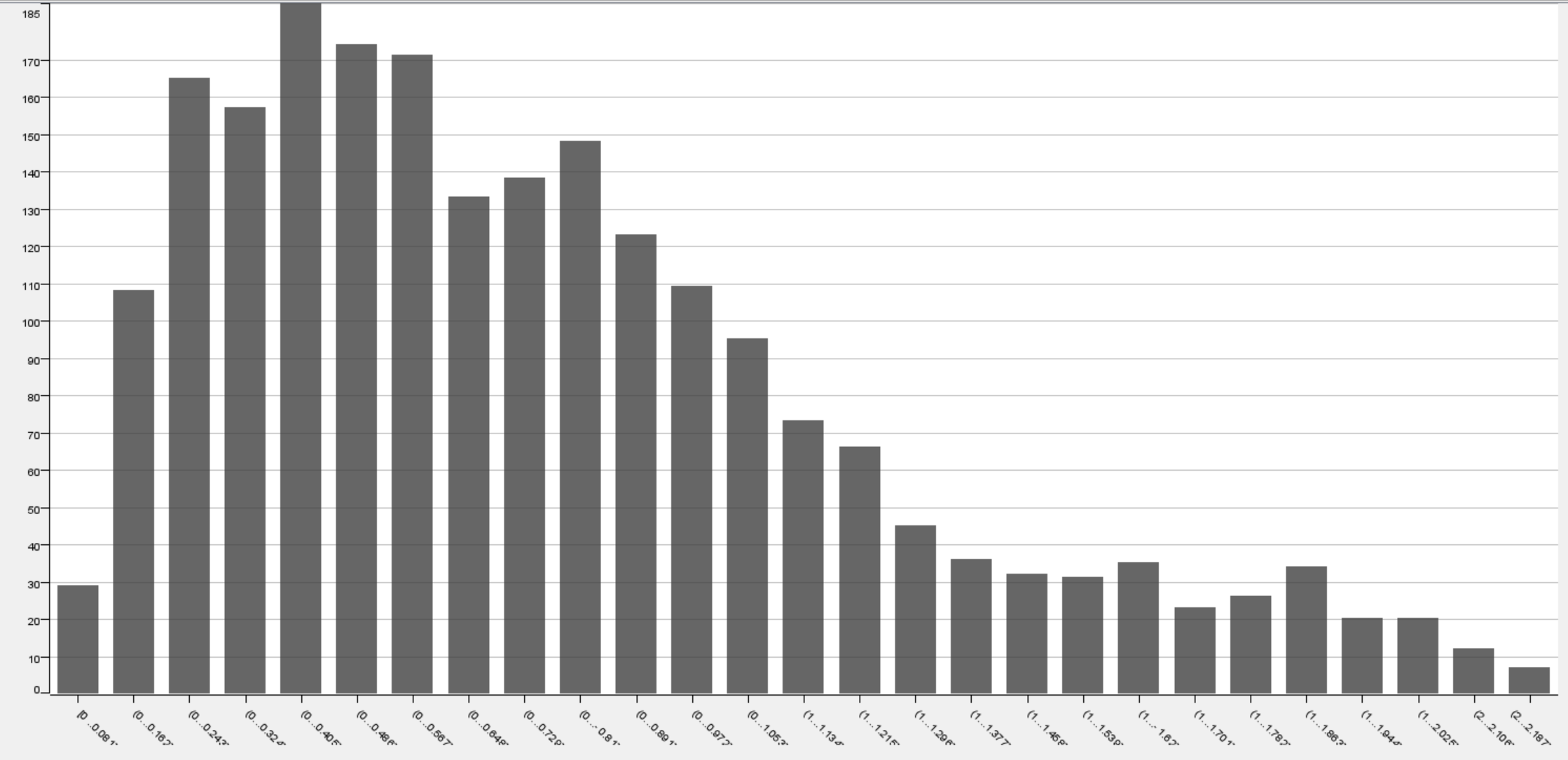
Description automatically generated

### 1.2.16 NumTransfers

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 1.131 |
| Minimum Value | 0 |
| Maximum Value | 75 |
| Standard Deviation | 3.315 |
| Variance | 10.988 |
| Range | 75 |
| Mode | 0.44 |
| Skewness | 16.488 |
| Unique Values | 258 |
| 25% Quantile | 0.39 |
| Median | 0.68 |
| 75% Quantile | 1.09 |
| Sum | 2668.98 |
| Kurtosis | 333.632 |
| Missing Values | 0 |

As NumTransfers is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 26) visualizes the row count frequency statistics for the number of labs. Binning has been used to best capture the shape of the distribution, however due to the extreme outliers a readable histogram cannot be created without modification. The most common (mode) number of transfers is 0.44. Additionally, Figure 26 depicts the strong positive skew of the data which indicates that a lower number of callouts is most common. Figure 26 is a histogram with an upper bound of so that readability of frequency is improved for the other values.

*Figure 26 Histogram for NumTransfers:*



### 1.2.17 NumChartEvents

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 510.966 |
| Minimum Value | 0 |
| Maximum Value | 4825 |
| Standard Deviation | 445.947 |
| Variance | 198,868.961 |
| Range | 4825 |
| Mode | 0 |
| Skewness | 2.02 |
| Unique Values | 2260 |
| 25% Quantile | 206.74 |
| Median | 409.9 |
| 75% Quantile | 680.53 |
| Sum | 1,205,368.07 |
| Kurtosis | 2.02 |
| Missing Values | 0 |

As NumChartEvents is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 27) visualizes the row count frequency statistics for the number of chart events for patients at the hospital. Binning has been used to best capture the shape of the distribution. The most common (mode) number of events is 0. Additionally, Figure 27 depicts the strong positive skew of the data which indicates that lower chart events are most common. Similarly, the below box plot (See Figure 28) clearly depicts that the majority of patients stay have less than 680.53 chart events.

*Figure 27 Histogram for NumChartEvents*

A graph of a number of people

Description automatically generated with medium confidence

*Figure 28 Box Plot for NumChartEvents*

A diagram of a diagram

Description automatically generated with medium confidence

### 1.2.18 ExpiredHospital

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | 0 |
| Unique Values | 2 |
| Missing Values | 0 |

As ExpiredHospital is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique Values, Missing Values and Frequency statistics. The below pie chart (See Figure 29) visualizes the frequency statistics and indicates the values (0, 1) that support the ExpiredHospital attribute. ‘1’ represents that the patient did expire in the hospital and ‘0’ represents that they did not. From the row count we find that 9.41% of patients expired and 90.59% of patients did not expire. From the frequency bar chart (See Figure 30) we see ‘0’ greatly outweights 1 with 2137 entries to 222 entries, hence the mode is ‘0’.

*Figure 29 Pie Chart for ExpiredHospital*

A gray circle with a number of percentages

Description automatically generated

*Figure 30 Bar Chart for ExpiredHospital*

A graph with a bar

Description automatically generated with medium confidence

### 1.2.19 TotalNumInteract

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mean | 616.366 |
| Minimum Value | 0 |
| Maximum Value | 10550 |
| Standard Deviation | 577.434 |
| Variance | 333,429.893 |
| Range | 10550 |
| Mode | 0 |
| Skewness | 4.133 |
| Unique Values | 2336 |
| 25% Quantile | 266.37 |
| Median | 493.62 |
| 75% Quantile | 786.89 |
| Sum | 1,454,007.59 |
| Kurtosis | 44.375 |
| Missing Values | 0 |

As TotalNumInteract is a ratio attribute, quantitative summary statistics of location and spread have been included. The below histogram (See Figure 31) visualizes the row count frequency statistics of patient interactions. Binning has been used to best capture the shape of the distribution. Figure 31 is a histogram but with an upper bound so that readability of frequency is improved for the other values. 0 is by far the most common (mode) number of. From the quantiles we also find that 75% of values lie between 0 and 786.89. This results in a strong positive skew to the age distribution as visualized by the histograms.

*Figure 31 Histogram for TotalNumInteract*

*A graph of a person with a beard

Description automatically generated with medium confidence*

### 1.2.20 Marital status

|  |  |
| --- | --- |
| **Statistics** | **Value** |
| Mode | MARRIED |
| Unique Values | 6 |
| Missing Values | 0 |

As Marital status is a nominal attribute, quantitative summary statistics of location and spread are meaningless. Hence, we include Mode, Unique values, Missing values and frequency statistics. The below pie chart (See Figure 32) and bar chart (See Figure 33) visualizes the row count frequency statistics and indicates the values (MARRIED, WIDOWED, SINGLE, LIFE PARTNER, DIVORCED, SEPARATED) that support the martial status attribute. The most common (mode) marital status is MARRIED which makes up 40.57% or 957 entries of all cases.

*Figure 32 Pie Chart for Marital status:*

A pie chart with numbers and a number of people

Description automatically generated

*Figure 33 Histogram for Marital status:*

**A graph of blue bars

Description automatically generated**

## 1.3 Exploration

### 1.3.1 Outliers

Row number 23 has an extreme outlier for NumChartevents of 4825 as demonstrated by figure 44. This value is so extreme it could be a mistake.

*Figure 44 Box Plot for NumChartEvents*

A screenshot of a graph

Description automatically generated

Similarly, in row number 2327 an extreme outlier of 2 NumCallouts was recorded as depicted in figure 45. This value is so extreme it could be a mistake.

*Figure 45 Box Plot for NumCallouts*

A graph with numbers and a box

Description automatically generated with medium confidence

The distribution of outliers is visualised below, interestingly there are no age outliers indicating that the hospital deals with all ages:

A screenshot of a number graph

Description automatically generated

### 1.3.2 Interesting attributes

NumChartEvents is an interesting attribute as it has a very strong correlation with TotalNumInteract (See figure 46) having a correlation value of 0.996. The almost one to one correlation in frequency between the two variables indicates that an interaction with a patient almost always requires a chart event. This can be explained by the need to update any changes or lack of changes to a patient condition for future reference and analysis. This also suggests that the hospital is successfully completing their due diligence and requirement to maintain record of patient care.

*Figure 46 Scatter Plot for TotalNumInteract vs NumChartEvents*

A graph with blue dots

Description automatically generated

Similarly, there is a strong linear correlation between NumDiagnosis and NumTransfers (See figure 47) with a correlation value of .759. This indicates that as a diagnosis is updated the necessary transfers are occurring. A zoomed in scatter plot which better illustrates this correlation is shown below.

*Figure 47 Scatter Plot for NumDiagnosis vs NumTransfers*

A graph with blue dots

Description automatically generated

We find the length of stay in days varies based on the type of insurance of the patient (See figure 48). Private insurance, which is typically more expensive than Medicaid and government insurance had a longer LOSdays. This could indicate that less fortunate patients are experiencing a worse standard of care and are being rushed out of the hospital.

*Figure 48 Scatter Plot for insurance vs LOSdays*

A screen shot of a graph

Description automatically generated

Curiously, as the number of transfers increases the length of stay decreases, despite a transfer involving processing that adds handling time for a patient (See figure 49). This could suggest that transfers are external between hospitals, while length of stay is for this specific facility.

*Figure 49 Scatter Plot for NumTransfers vs LOSdays*

A graph with numbers and dots

Description automatically generated

The rank correlation chart below illustrates the correlations between the attributes outlined in this report (See figure 50). Missing values were filled in to avoid filtering of these attributes. Interestingly there are very few strong negative correlations relative to the positive correlations in this data set. Rank correlation was chosen as it can handle categorical data automatically.

*Figure 50 Rank Correlation for attributes.*

A screenshot of a computer

Description automatically generated

The below table depicts the median values for ratio attributes based on whether that patient had expired or not. Median was chosen to avoid the impact of outliers. From this a future avenue of research would be to develop models that can predict patient outcomes for doctors etc. to gain a better understanding of the state of the patient and hence prioritise patients more at risk of death.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Median (expired)** | **Mean (not expired)** |
| Age | 65.5 | 57 |
| LOSdays | 5.9 | 6.33 |
| NumCallouts | 0 | 0.05 |
| NumDiagnosis | 1.925 | 1.37 |
| NumProcs | .58 | 0.38 |
| NumCPTevents | 1.725 | 0.91 |
| NumInput | 48.925 | 12.15 |
| NumLabs | 68.5 | 37.24 |
| NumMicroLabs | 1.53 | 0.47 |
| NumOutput | 11.72 | 4.9 |
| NumTransfers | 0.595 | 0.68 |
| NmChartEvents | 10001.885 | 388.24 |
| TotalNumInteract | 1444.025 | 467 |

### 1.3.3 Clusters

K means clustering was chosen in figure 51 for its superior silhouette coefficient compared to other methods, indicating valid clusters. Clusters are distinguished by colour and whether a patient expired in hospital is depicted by a square or circle for no and yes respectively. This was achieved by converting the ExpiredHospital attribute to a nominal attribute. There are 2 clusters to correlate to the 2 values for ExpiredHospital which we are trying to group the data by. It was found that this attribute beset matched the clustering. From figure 51 we see that patients who had over about 1000 interactions in hospital, regardless of their age, were more likely to expire. This is shown by the majority of circles lying above this point. A possible explanation for this is that patients requiring more interactions are generally sicker. This data could be used to predict patient expiration and hence predict hospital capacity based on current patients.

*Figure 51 Scatter plot of age vs TotalNumInteract clustered by expired.*

*A close-up of a sign

Description automatically generated*

*A screenshot of a computer

Description automatically generated*

A graph with blue and yellow squares

Description automatically generated

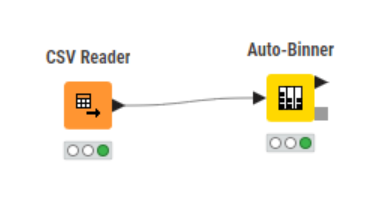
# 2. Data Preprocessing

## 2.1 Binning of age

Binning places attribute values into a designated ranges to better capture the overall grouping of data and reduce the number of attribute values.

### 2.1.1 Equiwidth Binning

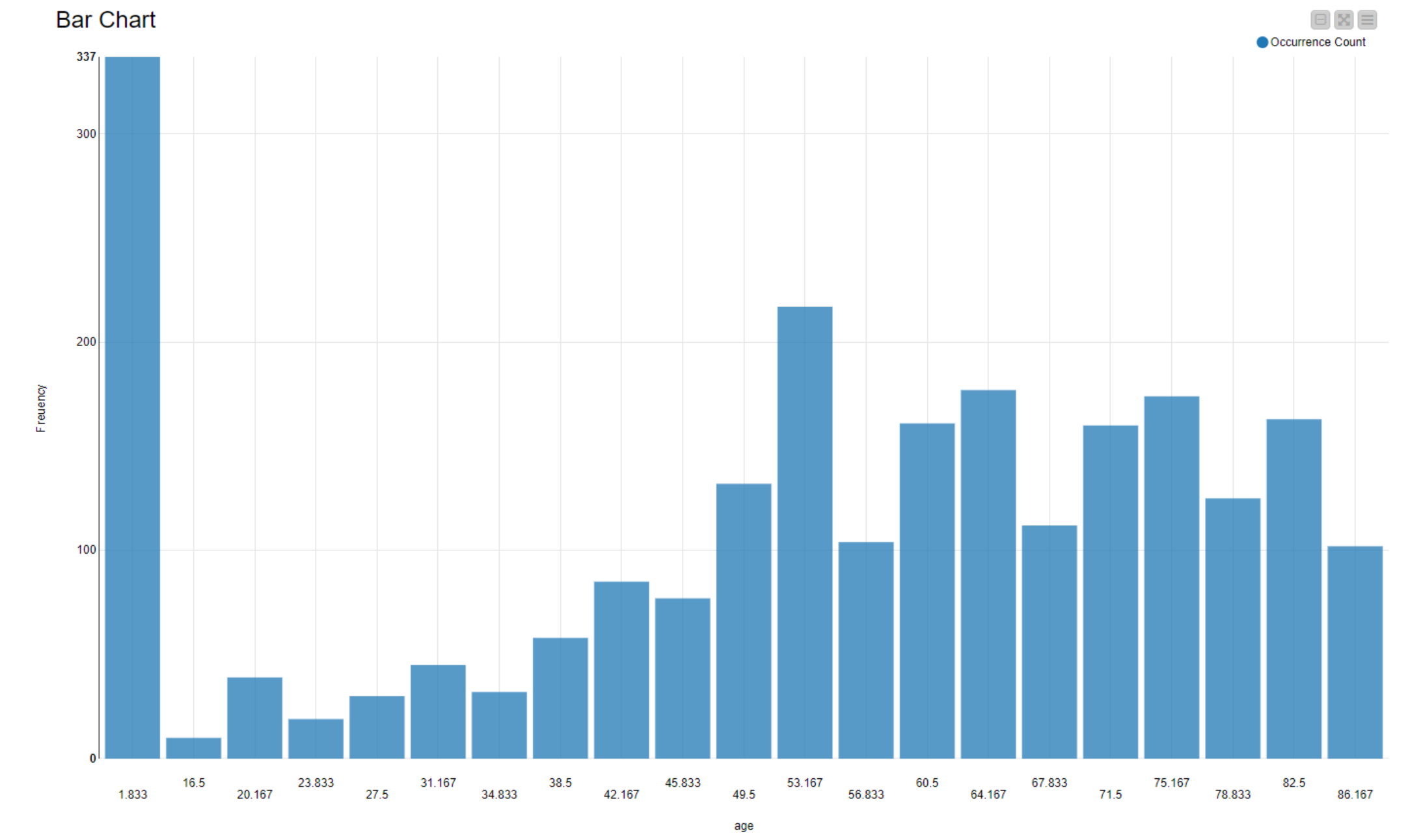
The Auto-Binner node is used to achieve this in KNIME as shown below:



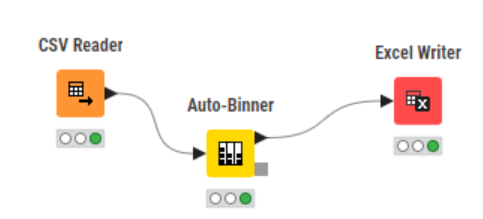
We Configure it with 23 bins as a lower number of bins was found to poorly convey the dominance of age 0 entries in the data set and any higher did not add value to the visibility of overall shape. Binning is labelled with the midpoint value of the bin. To visualise the binning and verify that the number of bins is optimal we send the output of the Auto-Binner into a bar chart node which is set to display the bins in alphabetical order using their row count frequency as shown below:

A diagram of a car

Description automatically generated



To export the binned data we apply the excel writer node:



This produces column B of the below table:

A screenshot of a table

Description automatically generated

### 2.1.2 Equidepth binning

This follows the same steps as equiwidepth binning, however the binning is set to frequency instead of width based. We set the number of bins to 6 as any higher and there was too much disparity in frequency between bins for equiwdepth, and any less and the shape of the distribution was not captured. The bar chart for this binning is depicted below:

A graph of blue bars

Description automatically generated with medium confidence

This produces column C of the below table:

A screenshot of a table

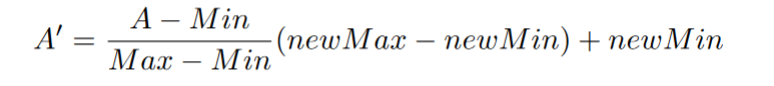
Description automatically generated

## 2.2 Normalisation of NumLabs

Attribute normalisation is the process of taking values that span a given range and representing them in a different range (Typically -1 to +1, or 0 to 1). This creates a common scale for numeric values. The goal of normalisation is to ensure that information is not lost in this process.

### 2.2.1 Min-max normalisation

Min-max normalisation involves a linear transformation from A to A’ as defined below. This preserves all relationships of data value without introducing bias:

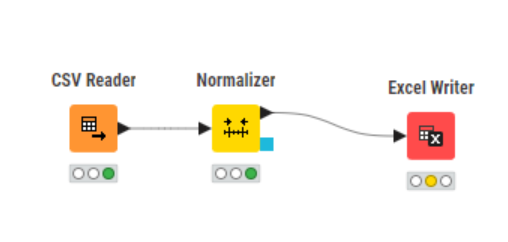


The Normalizer node is used to achieve this in KNIME as shown below:

A diagram of a normalized system

Description automatically generated with medium confidence

We configure it to use min-max normalisation with a range of 0.0 to 1.0. We then send the output of this node into the excel writer node:



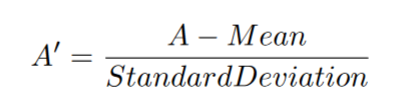
This produces column B of the below table:

A screenshot of a table

Description automatically generated

### 2.2.2 Z-score normalisation

Z-score normalises an attribute based on its mean and standard deviation. This creates the possibility for a wide variety of values as the mean and standard deviation depend on the data. The attributes are converted to their corresponding z-score which indicates how many standard deviations they are away from the mean, this is shown below and is effective when the actual minimums or maximums of the original data is not known, or whether there are outlier values that dominate min-max normalisation:



This follows the same steps as min-max normalisation however the normalisation method is set to z-score normalisation (Gaussian) in the normalizer node.

This produces column C of:

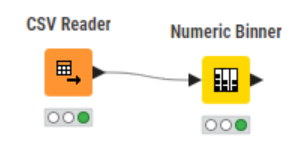
A screenshot of a table

Description automatically generated

## 2.3 Discretisation of LOSdays

Discretisation places the values of a continuous attribute into discrete intervals which reduces the number of values of the original data. The interval labels replace the actual data values. This can also be used to create a more conceptual description of data, e.g. replacing numeric age values by “young”, “middle aged” and “old”. In this case, we discretise LOSdays into shorter periods, medium periods, longer periods and very longer periods corresponding to 0-5, 5-15, 15-50 and 50+ days respectively. These ranges were chosen relative to the distribution of LOSdays data and a conceptual understanding of length of stay in a hospital environment. Additionally, the upper bound of these ranges are exclusive, whereas the lower limit is included in the range.

The Numeric Binner node is used as shown below and configured as previously outlined:



We then send the output of this node into the excel writer node:

A diagram of a number line

Description automatically generated with medium confidence

This produces:

A screenshot of a table

Description automatically generated

## 2.4 Binarisation of marital status

Binarisation is used to convert a categorical or continuous attribute into binary values. A new column is created for each possible value of a given attribute. In these new columns a 1 represents that the value of the instance for the given attribute is that of the column title, and a 0 represents that this is not the case.

In KNIME we use the One-to-Many node to achieve this:

A diagram of a computer

Description automatically generated with medium confidence

We then send the output of this node into the output of the excel writer node:

A diagram of a computer program

Description automatically generated with medium confidence

This produces:  
A table with numbers and letters

Description automatically generated

# 3. Summary

Some of the most important findings of this report include:

* AdmitDiagnosis is the only attribute with a non-zero number of missing values indicating the current record keeping is effective.
* NumLabs was the only attribute that resembled a normal distribution when ignoring the high frequency of 0 entries. This spike of 0 applied to most attributes, yet regardless of this feature there is positive skew to all attributes except age.
* One of the few strong negative correlations is between NumTransfers and LOSdays which needs further explanation as there are multiple hypothesis for this. One hypothesis is that transfers include external and off site trasnfers, in which case this distinction should be noted in the data. This would enable clustering analysis to be carried out based on whether a transfer was to or from an internal or external premise.
* The on average longer length of stay for patients with more expensive insurance could indicate a unfair distribution of care which should be explored to ensure all patients are being treated equally.
* Predictive relationships between attributes that affect LOSdays should be further explored which could provide a means to simulate patient turnover and future capacity based on the current patients to optimally organise scheduling.
* Clustering showed that after over about 1000 interactions a patient regardless of their age is more likely to expire in hospital. This data could be used to predict patient expiration and hence predict hospital capacity based on current patients.
* There is a tangible difference in median attribute values between patients who expire and patients who do not, this data could be used to predict higher risk patients who require more care.
* The attributes NumDiagnosis, NumProcs, NumCPTevents, NumInput, NumLabs, NumMicroLabs, NumOutput, NumTransfers, NumChartEvents and TotalNumInteract are all decimal values in the original data despite the quantities these attribute measure being conceptually discrete integer values. This could either be to issues in the dataset or normalisation that has already been done on the data. An explanation to this needs to be developed for these attribute values to be properly understood.
* The attributes age, NumProcs, NumCPTevents, NumInputs, NumMicroLabs, NumOutput, NumChartEvents, TotalNumINteract all have dominant modes of 0 which is not obviously captured in the histograms after binning.